International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 12, Number 2 (2016), pp. 209-214
© Research India Publications
http://www.ripublication.com

Mining Top K High Utility Itemsets and Frequent Patterns in one Phase

MS. SHILNA S[1] AND MS. NAVYA E K[2]

[1]Malabar Institute of Technology, Anjarakandy, India.
E-mail: shilna94@gmail.com

[2]Malabar Institute of Technology, Anjarakandy, India.
E-mail: Navya.06rimaan@gmail.com

Abstract

High-utility mining is an important data mining task with wide applications.Generally utility mining adopt a two-phase, candidate generation approach, that is,first find candidates of high utility patterns in the first phase, and again data is scaned and identify high utility patterns from the candidates in the second phase. The drawback is that the number of candidates can be huge, which is the scalability and efficiency bottleneck.To solve this drawback., finds high utility patterns in a single phase without generating candidates. d2HUP algorithm, for utility mining with the itemset share framework, finds high utility patterns in a one phase without generating candidates.To improve the efficiency of this algorithm and find top k high utility itemset ,TKO another algorithm in one phase is used.Along with this frequent pattern is also generated using CR Tree. High utility itemset mining is a research area of utility based descriptive data mining. Proposed system having applications in Business, Cross marketing in retail stores, online e-commerce management, Mobile commerce environment planning.

INTRODUCTION

Data mining is a powerful technology with high potential to help companies focus on the most important information in the collected data. It finds information within the data that queries and reports can’t effectively reveal. By performing data mining, knowledge, patterns, regularities or high-level information can be extracted from databases and viewed from different angles. This knowledge can be applied to decision making, process control, information management and query processing Mining top k high utility itemsets from transactional databases is an important data mining task, which refers to the find itemsets with high utilities (e.g. high profits).The utility of an itemset represents its importance, which can be measured in terms of weight, value, quantity or other information depending on the user specification. An itemset is said to be high utility itemset (HUI) if its utility is no less than a user-specified minimum utility threshold min-util. HUI mining is essential to many applications such as streaming analysis, market analysis, mobile computing and biomedicine. Finding interesting patterns has been an important data mining task, and has a variety of applications, for example, genome analysis, condition monitoring, cross marketing, and inventory prediction, where interestingness measures play an important role. With frequent pattern mining, a pattern is regarded as interesting if its occurrence frequency exceeds a specified threshold. For example, mining frequent patterns from a shopping transaction database refers to finding of sets of products that are frequently purchased together by customers. However, a user's interest may relate to many factors must not be expressed in terms of the occurrence frequency. For example, a supermarket manager may be interested in finding combinations of products with high profits, which relates to the unit profits and purchased quantities of products that are not considered in frequent pattern mining.

.

RELATED WORK

Guangzhu Yu et al. hybrid method[2], which is composed of a row enumeration algorithm (i.e., Inter-transaction) and a column enumeration algorithm which is Two-phase, to discover high utility itemsets from two directions: Two-phase finds short high utility itemsets from the bottom, while Inter- transaction finds long high utility itemsets from the top. In extension, optimization technique is adopted to improve the performance of computing the intersection of transactions. Experiments based on synthetic data show that the hybrid method achieves high performance in large high dimensional datasets.Disadvantages is hybrid method is not suited for datasets with only short patterns. The two-phase approach have scalability issue due to the huge number of candidates [1]. The challenge is that the number of candidates can be large, which is the scalability and efficiency bottleneck. Although a lot of effort has been made to reduce the number of candidates produced in the first phase, the challenge still persists when the raw data contains many large transactions or the minimum utility threshold is small. Such a huge number of candidates causes scalability problems in both the first phase and second phase, and consequently degrades the efficiency.

Vincent S. Tseng et al. proposed a new algorithm UP-Growth [3]. Mining high utility itemsets from a transactional database refers the finding of itemsets with high utility. Although a number of algorithms have been proposed in recent years, they incur the problem of producing a huge number of candidate itemsets for high utility itemsets. Such a large number of candidate itemsets decreases the mining performance in terms of execution time and space requirement. In this paper, we propose two algorithms, utility pattern growth (UP-Growth) and UP-Growth+, for mining high utility itemsets with a set of effective methods for pruning candidate itemsets. The information of high utility itemsets is maintained in a tree, utility pattern tree (UP-Tree) such that candidate itemsets can be generated efficiently with only two scans of database. The UP-Growth and UP-Growth+ performance is compared with other algorithms on many types of data sets. Results show that the proposed algorithms, UP- Growth+, reduce the number of candidates and outperform other algorithms in terms of runtime.Disadvantage is the algorithm generate large number of candidates.Which cause scalability issues since it a two phase algorithm.This algorithm may perform well only in some applications, they does not fin top-k high utility itemset mining and still suffer from the subtle problem of setting appropriate thresholds.

Charu C. Aggarwal et.al provide a detailed survey of frequent pattern mining algorithms[4]. A wide variety algorithms will be used starting from Apriori. Many algorithms such as Eclat, TreeProjection, and FP-growth. In data mining, frequent pattern mining (FPM) is one of the most investigated problems in terms of computational and algorithmic development. Many algorithms have been proposed in last two decades, to solve frequent pattern mining or some of its variants, and the interest in this problem still persists .Different frameworks have been defined for mining frequent patterns. The common one is the support-based framework, in which itemsets with frequency above a given threshold are found. However, such itemsets may not usually represent interesting positive correlations between items. Consequently, alternative measures for interestingness have been defined in the literature. One of the main reasons for the high interest in mining frequent pattern algorithms is due to the computational challenge of the task. Even for moderate sized dataset, the search space of FPM is inordinate, which is exponential to the length of ttransactions in the dataset. This naturally cause challenges for itemset generation, when the support levels are low.It is critical to perform the analysis in a space- and time-efficient way. The main focus of work was to find FPM algorithms with better computational efficiency.In fact, the execution tree of all algorithms is different in terms of order in which the patterns are explored, and also check whether the counting work done for different candidates is independent of one another.

EXISTING SYSTEM

High utility itemsets (HUIs) mining from databases is a data processing task that refers to the itemsets with HULS discovery. However, it should gift too several HUIs to users, that conjointly degrades the potency of the mining process.Setting min util appropriately is a difficult task for users. Discovering an appropriate minimum utility threshold by trial and error is a difficult process for users. If min util is set too low, too many high utility itemsets will be produced, which may cause the mining process to be inefficient. On the other hand, if min util is set too high, it is likely that no HUIs will be found.Existing algorithms for utility mining are inadequate on datasets with high dimensions or long patterns.They incur the problem of producing a large number of candidate itemsets for high utility itemsets. Such a large number of candidate itemsets decrease the mining performance in terms of execution time and space requirement.

Frequent itemset mining is a research area in the field of data mining with wide range of applications. Efficient algorithm to discover frequent itemset is essential in data mining. Apriori: uses a generate-and-test approach generates candidate itemsets and tests if they are frequent. Candidate itemsets generation is expensive(in both space and time) and Support counting is expensive.There are new algorithm on existing algorithm to solve data mining problem efficiently, and the interest in this problem still persists.

PROBLEM DEFINITION

Generally utility mining adopt a two-phase, candidate generation approach, that is,first find candidates of high utility patterns in the first phase, and then scan the raw data one more time to identify high utility patterns from the candidates in the second phase.Here d2HUP algorithm is used. D2HUP, namely Direct Discovery of High Utility Patterns, which is an combination of the depth-first search of the reverse set enumeration tree, the pruning techniques.D2HUP does not perform in all applications, they are not developed for top-k high utility itemset mining and still suffer from the subtle problem of setting appropriate thresholds. Finding an appropriate minimum utility threshold by trial and error is a difficult process for users.Generally we have many algorithm, but many are two phase.So need to scan the database twice.It is time consuming and requires more memory usage.These algorithm does not give top k high utility itemsets also.They incur the problem of producing a huge number of candidate itemsets for high utility itemsets. Such a large number of candidate itemsets decrease the mining performance in terms of execution time and space requirement. FP tree contains only frequent item. Sub-trees are not locally optimized. Ordering of items within paths from the root to leaves are ordered by global support. Children of a node are not sorted.

PROPOSED SYSTEM

In the proposed system Top-K High Utility Itemsets is mined by using a new algorithm named TKO in one phase.In TKO min util is not set.Instead a border minimum utility threshold is used. TKO can raise the min_util threshold as high as possible and as quickly as possible, and further reduce as much as possible the number of candidates and intermediate low utility itemsets produced in the mining process.A novel framework for top-k high utility itemset mining, where k is the desired number of HUIs to be mined. TKO (mining Top-K utility itemsets in One phase) are proposed for mining the complete set of top-k HUIs in databases without the need to specify the min_util threshold. The TKO algorithm uses a list-based structure named utility-list to store the utility infor- mation of itemsets in the database. It uses vertical data representation techniques to discover top-k HUIs in only one phase. It utilizes the basic search procedure of HUI-Miner and its utility-list structure. Whenever an itemset is generated by TKO, its utility is calculated by its utility-list without scanning the original database. We first describe a basic version of TKO named TKOBase and then the advanced version, which includes several strategies to increase its efficiency.Many different types of data structure and algorithm have been proposed to extract frequent pattern from a large given database. One of the fastest frequent pattern mining algorithm is the CR algorithm, Which can efficiently represent whole data structure over single scan of the database. We have proposed an efficient tree based structure CR Tree in terms of execution time and memory usage.Corelation Tree or CR Tree algorithms.. Even a huge database can be processed by CR Tree if out-of-date transactions are removed concurrently**.** CR Tree is better than FP tree. Contains all items in every transaction.

CONCLUSION

This work presents a algorithm of mining for high utility itemsets.Generally utility mining adopt a two-phase, candidate generation approach, that is, first find candidates of high utility patterns in the first phase, and then scan the raw data again to identify high utility patterns from the candidates in the second phase.It can discover top-k HUIs in only one phase. It utilizes the basic search procedure of HUI-Miner and its utility-list structure. Whenever an itemset is generated by TKO, its utility is calculated by its utility-list without scanning the original database.Along with this frequent patterns is also mined using a efficient algorithm CR Tree.CR Tree extends the idea of FP- Tree to improve storage compression and to allow mining frequent patterns without generation of candidate itemsets. The proposed algorithms enable mining frequent patterns with different supports without rebuilding the tree structure. Furthermore, the algorithms allow mining within one pass over the database as well as efficient insertion or deletion of transactions at any time.Customer segmenatation can be done based on the top k HUI's and frequent patterns found.This can be used to improve the business.

REFERENCES

[1] Junqiang Liu, Ke Wang, and Benjamin CM Fung. Mining high utility patterns in one phase without generating candidates. IEEE Transactions on Knowledge and Data Engineering, 28(5):1245–1257, 2016.

[2] GuangzhuYu, KeqingLi, andShihuangShao. Mininghighutilityitem-sets in large high dimensional data. In Proceedings of the 1st internationalconference on Forensic applications and techniques in telecommunications, information, and multimedia and workshop, page 47. ICST (Institute of Computer Sciences, Social-Informatics and Telecommunications Engineering), 2008.

[3] Vincent S Tseng, Bai-En Shie, Cheng-Wei Wu, and S Yu Philip.Efficient algorithms for mining high utility itemsets from transactional databases. IEEE transactions on knowledge and data engineering,

[4] Jiawei Han, JianPei, YiwenYin, and Runying Mao. Miningfrequentpat-terns without candidate generation: A frequent-pattern tree approach.Data mining and knowledge discovery, 8(1):53–87, 2004.

[5] Liu Yongmei and Guan Yong. Application in market basket researchbased on fp-growth algorithm. In Computer Science and InformationEngineering, 2009 WRI World Congress on, volume 4, pages 112–115, 2009.